Express Mailing Label No. EV 302 232 589 US

PATENT APPLICATION
Docket No.: 13768.417

UNITED STATES PATENT APPLICATION

of

Dmitriy Meyerzon

and

Kenji C. Obata

for

NORMALIZING DOCUMENT METADATA USING DIRECTORY SERVICES

NORMALIZING DOCUMENT METADATA USING DIRECTORY SERVICES

BACKGROUND OF THE INVENTION

1. The Field of the Invention

[001] This invention relates to systems, methods, and computer program products for improving computerized search functions by synchronizing document metadata using directory services.

2. Background and Relevant Art

[002] Computerized environments have increased the efficiency by which people perform a wide variety of tasks. For example, computers and computer networks have vastly improved the speed and capabilities by which people communicate ideas to each other. Computerized systems also provide people with enhanced tools for fixing varyingly complex thoughts into an easily accessible medium, which provide far more options than typewriters, pens, pencils, and notepads. Thus, computerized systems greatly enhance information access, and authoring power. In these regards, the advantages of computerized systems are well known.

[003] With regard to information creation, one can author (or create) information as simply as by typing one or two basic text paragraphs in a document that the author may wish to send to another over electronic mail (E-Mail). In other cases, one or many authors may generate thousands of pages in a word processing document, where the word processing document may include several spoken languages, may contain several graphics and other multi-media content, and may comprise a wide variety of electronic formats. In any case, common electronic tools such as a word processor, a web page

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

creator, a text form, and so on help authors affix huge amounts of information into a wide variety of accessible electronic media.

[004] Computerized systems have also enhanced the speed and ability to locate and access this information created by others. Information is accessed and distributed using any one of a number of different techniques and applications, including electronic mail; distributed networks including the Intranet and corporate intranets; database storage and access systems; and the like. However, the overwhelming amount of data and information that is accessible has given rise to problems. In particular, the ability to specifically locate a relevant piece (or pieces) of information, such as a document, from a large and distributed database of information, such as the Internet or a large corporate intranet, has proved to be increasingly difficult.

[005] To address this problem, various types of search tools, sometimes referred to as "search engines," have been developed. While any one of a different number of techniques and search algorithms are used, in general users of a search engine typically enter one or more search terms, and search results corresponding to those terms are returned by the search engine. In a typical implementation, a person may visit an Internet or local webpage that employs a functional text-input box. The person enters one or more search terms into the text box and the search engine may return to the person one or more related documents, depending on the specificity and nature of the search request.

[006] Search engine implementations vary in complexity and capability. A very simple search engine, for example, may only search for exact spellings of certain words within an opened document. Thus, if a user were to type the misspelled word "medixcal" into a search box, the simple search engine will not likely return any results,

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

or point the user to any meaningful point in the document, unless that exact misspelling has been made within the document. A more complex search engine, however, may allow a user to search millions upon millions of documents based on a wide variety of criteria, even allowing the user to add detailed restrictions, all the while compensating for misspellings. For example, a complex search engine may allow a user to search millions of documents on a local or wide area network for the terms "Fyre Engine" + "fire pole" + "Dennis Finch", with the restriction that all results must be in English, and that the resultant web page be created after the year 1998. In some cases, the search engine may even correct the spelling of "Fyre Engine" to "Fire Engine", prior to executing the search.

[007] Figure 1 illustrates a prior art depiction of one example of an implementation of a search engine. In this example, the search engine algorithm first obtains several documents (or any analogous discrete unit of information) 105, 110 into a database such as index service 100. In one approach, a user may select and enter documents 105 and 110 manually into the index service 100 to be processed. Alternatively, the index service (or related search service) may have a function that automatically locates and obtains documents. This function is sometimes referred to as "crawling," where the service continually "crawls" across multiple documents on the network by following document reference links within certain documents, and then processing each document as found. The index service 100 processes the documents by identifying key words or general text in the documents 105, 110, and then creating an inverted list 120 (more generally, an "index").

[008] An exemplary inverted list can be one or more electronic reference documents having a column containing a list of key words, a column containing one more

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

documents containing the key word, a column for the number of occurrences of that key word in the respective document, and a column with an address for each associated document. For the purposes of illustration, however, a more simple inverted list 120 is shown having a column of words (A, B, C, etc.), and a column indicating in which document those words can be found.

[009] When a user enters in one or more search terms (e.g., "Request for A" 132), a typical search engine 130 will employ an algorithm that first finds the one or more terms among the key words in the inverted list 120, and then weighs the resultant documents associated with any found words in the list. The search engine can then return one or more of the associated document references as results 136 to the user, depending on how the search algorithm is configured (i.e., documents having the most occurrences of the word), or depending on any restrictions the user places on the search (i.e., requiring an exact phrase match). Consequently, search engines can be quite useful for locating and accessing information contained within, for example, a distributed network environment.

[010] Search engines such as the foregoing, however, tend to have certain limitations. Since the typical such search engine relies on a generated index to locate documents, the relevance of a give search result is highly dependent on the document content that is used to ultimately construct the index. For example, a document containing only the words "whale," "fish," "ocean," and "ferry" would not be found by some search engines if a user entered the terms "orca," "tuna," "sea," and "transport." This is because, in general, search engines of the type described do not generate alternate word relationships when building an inverted list. While this type of search engine may

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

provide automatic spell-checking of search terms, they do not automatically search word variants, synonyms, and homonyms, unless the user specifically enters them.

[011] In addition, there are other problems that can complicate the amount and quality of data that a search engine can return to a user seeking information. For example, a large organization may have thousands upon thousands of internal documents on various topics posted on various servers on the local or wide area network. While each posted document may contain different metadata corresponding to metadata concepts (i.e., document identifiers) such as author, date created, size, title, etc., each document may be created with different programs that identify metadata properties differently, or describe the underlying data differently. For example, one document might include author metadata as: "Author = 'Heather F. Pettingill'" while another document's metadata might designate the author as: "By = 'H. Pettingill'" while yet another document might contain no author metadata and merely include the phrase "H. F. Pettingill" centered at the top of the first page. Thus, if an index were created that includes author metadata, a subsequent search may not locate some of these documents if a search were performed for the author "H. F. Pettingill."

[012] Even if the metadata format is standardized within the organization, the underlying data values that employees may use to classify documents within a general concept in the organization can often undergo several changes. For example, employees may refer to several documents under the classification of "Product Design" one year, and then "Manufacturing Policies" the next year when referring to the same general concept or classification. Similarly, a person's name or contact information may change several times over the course of their employment (i.e., due to name

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

changes, email alias changes, new email domain name, new preferences, new office, new workgroup, etc.).

[013] As such, this can degrade the effectiveness of searches, for example, for all documents authored by "Heather Pettingill," or for all documents discussing product release policies over the last three to five years. For example, with specific reference to Figure 1, if there were no direct correlation made between the values A and X on inverted list 120 such that A=X (e.g., "Heather Pettingill" = "Heather Martin" due to a marital name change), a normal search for A or X would only return "Doc" or "Doc2" (but not both) as a result 136. Typically, the only way the search engine might return both documents is if the user searched for both terms based on prior knowledge of the term correlation. Of course, this approach is limited by the fact that the user may not realize the correlation, though the user wishes to have all documents authored by the person in question.

[014] Accordingly, there is a need for more robust systems, methods, and computer program products that relate the types of information available to a search engine so that more accurate search results can be obtained, without requiring a user to iteratively search several variations of the same terms and phrases. In addition, there is a need for robust systems, methods, and computer program products that allow users to search returned results for additional relationships, such as by metadata concepts, or classification data.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

BRIEF SUMMARY OF THE INVENTION

[015] The present invention solves one or more of the foregoing problems in the prior art by introducing systems, methods, and computer program products for normalizing data -- such as metadata -- for use in a searching algorithm. In one embodiment, a filtering or indexing service retrieves alias information from one or more accessible directory services. The inventive method retrieves this information to normalize such things as contact information, classifications, metadata references, and so on. For example, if the accessed directory service is a personal directory service, the directory service might include a database of all versions of a name by which a person has been (or currently is) recognized. Each version of the person's name can constitute an alias (or, alternate identity).

[016] The directory service may also include various other forms of the person's contact information such as email aliases, prior and current workgroups, office locations and the like. In the case of personal contact information, each of the person's name aliases would then constitute a normalize-able identifier for the same person. The indexing or filtering service may also refer to other forms of information in the directory service as a type of class such that every alias of a person's name might constitute an "Authorship" class, whereas other types of alternate information might constitute a "Workgroup" class, etc. Each class may also be identifiable by one or more aliases (or, alternate identifiers).

[017] Once such information has been normalized (or, "aliased") from the directory service, a gathering service receives one or more documents as inputs, or crawls through various document links on a network, and identifies information segments in the documents as relating to the normalized data. The gathering service then stores

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

these relationships in a way that is accessible to a search engine. Thus, when a user enters a search term into a search engine input box, the search engine consults the alias and/or classification databases as appropriate, and returns a more robust result.

[018] Thus, a user can enter a person's name in one form and get a result from the search engine of all the documents related to that person's name, in whatever form the person has used, without requiring user knowledge of each of the forms. In addition, a user can make an initial search (or search returned results) to find additional property data from the search results that may relate to one or more aliased classes. For example, the user may enter a search of all documents created for the new topic name "Product Design", and the search engine can return all the documents related to "Product Design", without missing documents with a different topic name (e.g., "Manufacturing Specs").

[019] Additional features and advantages of the invention are set forth in the description which follows, and in part will be obvious from the description, or may be learned by the practice of the invention. The features and advantages of the invention may be realized and obtained by means of the instruments and combinations particularly pointed out in the appended claims. These and other features of the present invention will become more fully apparent from the following description and appended claims, or may be learned by the practice of the invention as set forth hereinafter.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

BRIEF DESCRIPTION OF THE DRAWINGS

[020] In order to describe the manner in which the above-recited and other advantages and features of the invention cam be obtained, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of the invention and are not therefore to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings in which:

[021] Figure 1 illustrates a prior art depiction of an indexing and searching service.

[022] Figure 2 illustrates one embodiment of representing document data objects and identifiers in accordance with the present invention.

[023] Figure 3 illustrates one embodiment of an overview block diagram of inputs and outputs of a gatherer data structure.

[024] Figure 4 illustrates a flow chart that may be used to implement a gathering process in accordance with the present invention.

[025] Figure 5 illustrates a flow chart of acts and functional steps for practicing one embodiment of the present invention.

[026] Figure 6 illustrates a suitable computing environment that may be used to practice the present invention.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[027] The present invention extends to both methods and systems for normalizing document metadata using directory services. Disclosed embodiments may comprise a special purpose or general-purpose computer including various computer hardware implementations, as discussed in greater detail below.

[028] Embodiments within the scope of the present invention also include computer-readable media for carrying or having computer-executable instructions or data structures stored thereon. Such computer-readable media can be any available media that can be accessed by a general purpose or special purpose computer. By way of example, and not limitation, such computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to carry or store desired program code means in the form of computer-executable instructions or data structures and which can be accessed by a general purpose or special purpose computer.

[029] When information is transferred or provided over a network or another communications connection (either hardwired, wireless, or a combination of hardwired or wireless) to a computer, the computer properly views the connection as a computer-readable medium. Thus, any such connection is properly termed a computer-readable medium. Combinations of the above should also be included within the scope of computer-readable media. Computer-executable instructions comprise, for example, instructions and data which cause a general purpose computer, special purpose computer, or special purpose processing device to perform a certain function or group of functions.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

[030] Reference is next made to Figure 2, which illustrates an exemplary representation of document data objects and identifiers that might be used in connection with an embodiment of the present invention. Illustrated is an electronic document 200, which may be of a wide variety of formats. In the example, document 200 can be broken down into at least two component parts: a content portion 205 and a metadata portion 210. In this example, the content portion 205 is the data entered by a user creating the document, e.g., the text of a word processing document. The contents 205 can be broken down into sections that represent size portions of the document, for example by page, or by words, phrases, and letters in the case of a text document. Here, the document 200 has multiple "content segments" (also referred to as "chunks") such as a name "Aaron J" 202, "Surgery" 206, and "Pediatric" 208. Each of these content segments can be representative of a particular property of the document.

[031] The other component part of the document, referred to here as metadata 210, represents specific properties or characteristics associated with the document. Metadata may be entered by a user when creating a document, or may be entered through default settings in an application program. For example, one might discover metadata by opening a "properties" portion of an opened document, or by some similar method. Thus, metadata 210 are meant to give information that can help characterize the contents in some way. By way of example, the metadata 210 in Figure 2 include metadata concepts (or "segments"), such as an "Author" segment 210, a file "Size" segment 214, and a "Keywords" segment 216. The metadata concepts are like classifications in the sense that they can be more generalized concepts of the more specific, underlying data content 205. Thus, a class is appropriately defined as a set of documents, and a label (e.g., all documents "Authored" by person A).

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

[032] For example, the "Author" segment 210 indicates that "AEJ" and "NDS" are authors of the document 200. It could be that both "AEJ" and "NDS" manually entered this information when initially creating the document, or that an application may have merely tracked this information as both "AEJ" and "NDS" entered information into the document from different locations. The "Size" segment 214 indicates that the size of the document 200 is 52 Kb. The "Size" segment 214 could be a value that is automatically updated when the document is saved, for example. The "Keywords" segment 214 includes keyword data that represents, for example, the general concepts addressed by the content of the document. Here, keywords segment 214 shows that the document 200 content is related to the terms "Science, Med, Osteo, Orthopedic, Surgery, and Pediatric." Again, these keywords could be manually entered, or could be automatically extracted from the content of the document itself.

[033] Figure 2 also shows an exemplary directory service 230, as can be used in connection with the example implementation. In the example, the directory service may contain one or more entries (such as is denoted here at 232 and 234), that each contain one or more data fields. For example, the entry 232 is for "Aaron Jones." That entry includes data fields for additional name aliases "AAJ," "AEJ," as well as different email addresses at different Internet Service Providers (ISPs). In addition, there is a secondary identifier "ID:ORTHOPEDICS." Similarly, there is an entry 234 for Nathan Smith, which includes data fields for name aliases "NDS" and "NAS," as well as different email addresses at different ISPs. This entry also includes a data field for a secondary identifier, shown here as "ID:PEDIATRIC SURGERY." It will be appreciated that these types of data fields are merely examples, and that any one of a different number and/or type of data field may included in a particular entry. Also, it

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

will be appreciated that a directory service may be as complex as a centralized, relational database in an organization, or may be as simple as a single text document relating one or more terms to one or more other corresponding alternative terms.

[034] As is further shown in Figure 2, each of the data field indicia (and other fields not shown) contained in the relative entries 232, 234 can be parsed by a gatherer function, denoted at 260, to provide an individualized association database 240 (e.g., an "alias" database). In the example shown, the association database 240 contains information from the parsed directory service 230. More particularly, in this example each contact entry 232, 234 is associated with one or more classes, denoted here at 246 and 256. As will be explained further, classes may be based on a general metadata concept, or may comprise a wide variety of other types of designations. In the example shown, one type of class in the database 240 could be authorship, and each contact entry 232, 234 is associated with one or more documents 200. The association database 240 can be configured to relate a wide variety of data from the directory service, and may be combined with a searchable index (e.g., the inverted index 120 in Figure 1) to provide an enriched inverted index, as will be discussed in more detail.

[035] Figure 3 illustrates one embodiment of the methodology by which the gatherer function, here designated at module 300, generates a searchable index 360 based on a variety of inputs. A gatherer 300 may be a functional executable program, or may be a plug-in module to a more complicated overall executable program such as a search engine. In either case, a gatherer 300 references one or more directory services (denoted in the example as 340 and 350) to ascertain a collection of aliases for one or more terms, as described in Figure 2. The gatherer 300 may reference the directory services as the gatherer 300 processes various documents on the network, or the

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

gatherer may receive a list of values (e.g., aliases) from the directory service in the form of a file that the gatherer recognizes as a definition set of aliases for different terms. The provided aliases allow the gatherer to have a "normalized" frame of reference for associating a network document with an entity or class. In other words, the gatherer 300 can now recognize that several word variants mean the same thing.

[036] To process documents, the gatherer 300 may first receive a document 305 as an input in some cases, such as a user entering certain document metadata before the user saves the document 305 on a network database. Or, the gatherer 300 may "crawl" throughout documents on a local or wide area network. As noted, crawling refers to a program module following links between various documents on a network in order to find and process additional documents. A gatherer 300 may crawl a network on its own initiative based on preset parameters, or a user may "seed" the gatherer 300 with a first network-based Uniform Resource Identifier (URI) 302. Once the gatherer 300 receives an input seed, the gatherer 300 may then follow any links or references contained within the document located at the first URI 302, following a trail through a second URI 315, a third URI 320, and a fourth URI 325 as it processes the respective documents.

[037] The gatherer 300 processes the respective document by breaking the encountered document into one or more segments. In some cases, the gatherer 300 may first need to decode the document it is processing into a readable format. Once the gatherer 300 can read the format, and has divided the document into segments, the gatherer 300 attempts to identify portions of the segments based on the normalized alias information the gatherer 300 has referenced from the one or more directory services (e.g., 340, 350). For example, the segments may be some form of defined metadata

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

concepts (e.g., "Author," "Type," "Workgroup," etc.), or some type of non-defined text (e.g., underlined section headings).

[038] For example, the gatherer 300 might identify URI_2 315 as having an "Author" value of "NDS", and a "Keyword" value of "pediatrics". Alternatively, the gatherer 300 may find no such formal metadata concepts in URI_4 325, but notice that the name "AJONES@ISP.NET" is centered at the top of the document, and so assign that document an "Author" value of "AJONES@ISP.NET." After consulting the alias references, the gatherer 300 could then identify that "NDS" = "Nathan Smith" 234, and that "AJONES@ISP.NET" = "Aaron Jones" 232, along with other values, and then place those associations for the documents found at the respective URIs in a searchable index 360.

[039] An exemplary searchable index 360 may have a defined set of classes 370 such as an "Author" class, a "Workgroup" class, and may further include a set of class aliases by which each class might be known in the organization, as well as a list of documents or entities that could belong to those classes. The searchable index 360 may also include a collection of entities 380 that relate to each of the classes. Entities may represent persons (e.g., a collection of aliases that identify the person) or class members, are used to establish relationships across classes, and are used for alias mapping. Class members are usually documents, or indexable entities having some property (i.e., metadata) that identifies them with a certain class. A class is then a "set" (i.e., a collection of objects that can be identified as belonging to the set) of these class members, and can be represented as a "set" in the mathematical sense. Thus, as in the example above, contact entry 232 for "Aaron Jones" may be represented as an entity in

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

a searchable index, where the entity may include several name and group aliases, and may also include other types of values relating to one or more classes.

[040] Figure 4 illustrates a flow chart showing the functional steps that may be used to implement one example implementation of the gathering process, particularly in the context of recognizing classes. For purposes of illustration, the gatherer is assumed to have already received one or more values from a directory service such that the gatherer has a reference for one or more aliases and or one or more classes. Continuing with Figure 4, a document 400 is received by a gatherer (e.g., gatherer 300), and then filtered 405 into one or more documents segments. Once the document has been filtered 405 into one or more segments, the gatherer processes the segment 407 by determining if the segment represents the end of the document 410. For example, the gatherer may look for a document-ending marker such as an End Of File (EOF) designator.

[041] If the segment does not signal the end of the document 412, then the gatherer processes the segment further to determine if the segment matches one of the class types (e.g, "By AEJ"). If the segment appears to be a valid class type 425, then the gatherer will seek to identify 430 whether there is a known class alias for the apparent class type (e.g., "By=Author=From=Written_By"). If there is no known class alias for the segment so that "By" represents any specific meaning, the gatherer will move to the next segment 450 for processing. If on the other hand, there is 445 an alias identifying the class type (e.g., "By" is recognized as an "Author" metadata concept), then the gatherer associates 460 the segment with the appropriate class by mapping the segment value to a preferred class label, and looks to see if any metadata associated with the class type are also aliases (i.e., "normalized" for a standard value – "AEJ = 'Aaron E. Jones'") so that the class type property may be called for the normalized metadata (or

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

alias for the normalized metadata) and vice versa. Having associated 460 the segment with any available classes and/or corresponding metadata, the gatherer adds the segment to a searchable index (e.g., 360) by association with the normalized value, and then proceeds to get the next segment 450.

[042] If the next segment 450 resembles an end of document designator 410, the gatherer determines if there is a next document 470 to filter into segments. For example, the gatherer may detect that there is an embedded URI for a next document (e.g., "crawling"), or the gatherer may detect a next document as input received in queue from another input method (e.g., a file list). If the gatherer detects a next document, 495, then the gather proceeds to filter the next document into segments, and continues the previously-described processing path. Alternatively, if the gatherer detects no new document, the gatherer ceases processing documents 480.

[043] Example embodiments of the present invention also may be described in terms of methods comprising functional steps and/or non-functional acts. The following is a description of acts and steps that may be performed in practicing one exemplary embodiment. Usually, functional steps describe the invention in terms of results that are accomplished, whereas non-functional acts describe more specific actions for achieving a particular result. Although the functional steps and non functional acts may be described or claimed in a particular order, the present invention is not necessarily limited to any particular ordering or combination of acts and/or steps.

[044] Figure 5 illustrates a flow chart of acts and functional steps for practicing one embodiment of the present invention, and will be described with reference to the foregoing figures. As shown, the inventive method may begin by performing the act of first receiving 500 a document 205 containing document data. As described, the

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

document may contain specific references to metadata concepts and may have as values the underlying metadata corresponding to the metadata concepts. Having received the document 500 as input (or having crawled to the document via a URI), the inventive method may parse 510 the document into one or more document segments. This may be done by, for example, filtering the document 405 into document segments as described in Figure 4. These filtered segments may as simple as data blocks having text values 202, 206, and 208, or may be more complicated in the sense of containing text and textual relationship values to a corresponding class or metadata concept, such as data blocks 210, 214, and 216.

[045] Thereafter, the inventive method performs the functional step 520 of normalizing document metadata used as a reference by a search engine. This step can include normalizing document metadata used as a reference by a search engine by maintaining one or more relationships between a search term and an alternate search term, a search term property or alternative search term property. Functional step 520 may be performed by the specific non-functional act of identifying 530 at least one of the one or more document segments as an alias; and by a non-functional act of associating 540 the received document with the document alias.

[046] Act 530 may be performed by identifying at least one of the one or more document segments as an alias for a document datum found in an alias directory service. Thus, for example, the inventive method may process a directory service and then store a series of aliases that refer to a main text value, such as in the case of "Aaron E. Jones" (i.e., contact entry 232) being recognized by several monikers such as "Aaron Jones", "AAJ", "AEJ, "AEJONES@ISP.NET", etc. Similarly, the inventive method may additionally or alternatively store class type values so that, for example, a search

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

for all documents authored by a person can be found without regard to the various aliases the person has used over time. A document datum found in an alias directory might include, for example, the name "NDS" as an alias for the name "Nathan Smith" in contact entry 234, and that "Nathan Smith" is further associated (via an associations database 240) as an "Author" (e.g., class 256) of a particular document (e.g., document 200).

[047] Act 540 may be performed by associating the received document with the document alias so that, upon request for the document datum through a search engine, the received document is returned to the requester by association of the document datum with the alias. Thus, for example, if a person enters the term "Nathan Smith" into a search engine input box and then executes the search, the search engine will look for any occurrence of the term "Nathan Smith". The search engine will also look for all documents containing the associated aliases (e.g., terms "NDS", and "NAS", from contact entry 242 of Figure 2), since the term "Nathan Smith," and each of the received aliases have been normalized.

[048] In addition, the concept of normalizing the class type (e.g. "Author") can be applied to train a statistical model used by the gatherer based on each set of documents (tagged with all alias forms) belonging to the respective class. That is, embodiments of the present invention may progressively modify its understanding of search terms to interrelate and associate concepts and metadata more correctly, or to "fine-tune" how the gatherer 300 relates documents and document data. For example, after crawling various documents on a network, the gatherer 300 may be trained to correctly discover other related documents to the given author, and other people related to the author, or to identify other concepts related to documents associated with the author, and so on.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

Accordingly, a statistical model can be trained for each class, so that there are as many statistical models as there are classes (e.g., in the case of authors, there would be a model for each author). Since a statistical model is usually represented as a set of key terms (i.e., keywords) with associated weights (i.e., importance), it is important to identify the set of documents that belong to the class correctly. Using aliases, as described herein, is one way of identifying the correct set of documents, and, ultimately, training of the statistical model.

[049] Those skilled in the art will appreciate that the invention may be practiced in network computing environments with many types of computer system configurations, including personal computers, hand-held devices, multi-processor systems, microprocessor-based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like. The invention may also be practiced in distributed computing environments where local and remote processing devices perform tasks and are linked (either by hardwired links, wireless links, or by a combination of hardwired or wireless links) through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[050] Figure 6 and the following discussion are intended to provide a brief, general description of a suitable computing environment in which the invention may be implemented. Although not required, the invention will be described in the general context of computer-executable instructions, such as program modules, being executed by computers in network environments. Generally, program modules include routines, programs, objects, components, data structures, etc. that performs particular tasks or implement particular abstract data types. Computer-executable instructions, associated

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

data structures, and program modules represent examples of the program code means for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps.

[051] With reference to Figure 6, an exemplary system for implementing the invention includes a general-purpose computing device in the form of a conventional computer 620, including a processing unit 621, a system memory 622, and a system bus 623 that couples various system components including the system memory 622 to the processing unit 621. The system bus 623 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory includes read only memory (ROM) 624 and random access memory (RAM) 625. A basic input/output system (BIOS) 626, containing the basic routines that help transfer information between elements within the computer 620, such as during start-up, may be stored in ROM 624.

[052] The computer 620 may also include a magnetic hard disk drive 627 for reading from and writing to a magnetic hard disk 639, a magnetic disc drive 628 for reading from or writing to a removable magnetic disk 629, and an optical disc drive 630 for reading from or writing to removable optical disc 631 such as a CD ROM or other optical media. The magnetic hard disk drive 627, magnetic disk drive 628, and optical disc drive 630 are connected to the system bus 623 by a hard disk drive interface 632, a magnetic disk drive-interface 633, and an optical drive interface 634, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-executable instructions, data structures, program modules and other data for the computer 620. Although the exemplary environment described herein employs a

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

magnetic hard disk 639, a removable magnetic disk 629 and a removable optical disc 631, other types of computer readable media for storing data can be used, including magnetic cassettes, flash memory cards, digital versatile disks, Bernoulli cartridges, RAMs, ROMs, and the like.

[053] Program code means comprising one or more program modules may be stored on the hard disk 639, magnetic disk 629, optical disc 631, ROM 624 or RAM 625, including an operating system 635, one or more application programs 636, other program modules 637, and program data 638. A user may enter commands and information into the computer 620 through keyboard 640, pointing device 642, or other input devices (not shown), such as a microphone, joy stick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 621 through a serial port interface 646 coupled to system bus 623. Alternatively, the input devices may be connected by other interfaces, such as a parallel port, a game port or a universal serial bus (USB). A monitor 647 or another display device is also connected to system bus 623 via an interface, such as video adapter 648. In addition to the monitor, personal computers typically include other peripheral output devices (not shown), such as speakers and printers.

[054] The computer 620 may operate in a networked environment using logical connections to one or more remote computers, such as remote computers 649a and 649b. Remote computers 649a and 649b may each be another personal computer, a server, a router, a network PC, a peer device or other common network node, and typically include many or all of the elements described above relative to the computer 620, although only memory storage devices 650a and 650b and their associated application programs 636a and 636b have been illustrated in Figure 6. The logical

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

connections depicted in Figure 6 include a local area network (LAN) 651 and a wide area network (WAN) 652 that are presented here by way of example and not limitation. Such networking environments are commonplace in office-wide or enterprise-wide computer networks, intranets and the Internet.

[055] When used in a LAN networking environment, the computer 620 is connected to the local network 651 through a network interface or adapter 653. When used in a WAN networking environment, the computer 620 may include a modem 654, a wireless link, or other means for establishing communications over the wide area network 652, such as the Internet. The modem 654, which may be internal or external, is connected to the system bus 623 via the serial port interface 646. In a networked environment, program modules depicted relative to the computer 620, or portions thereof, may be stored in the remote memory storage device. It will be appreciated that the network connections shown are exemplary and other means of establishing communications over wide area network 652 may be used.

[056] The present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive. The scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes that come within the meaning and range of equivalency of the claims are to be embraced within their scope.

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111